

Inferring Mobility Measures from GPS Traces with Missing Data

Ian Barnett

Jukka-Pekka Onnela

Harvard University, Boston, USA.

E-mail: onnela@hsph.harvard.edu

Summary. With increasing availability of smartphones with GPS capabilities, large-scale studies relating individual-level mobility patterns to a wide variety of patient-centered outcomes, from mood disorders to surgical recovery, are becoming a reality. Similar past studies have been small in scale and have provided wearable GPS devices to subjects. These devices typically collect mobility traces continuously without significant gaps in the data, and consequently the problem of data missingness was safely ignored. Leveraging subjects' own smartphones makes it possible to scale up and extend the duration of these types of studies, but at the same time introduces a substantial challenge: to preserve a smartphone's battery, GPS can be active only for a small portion of the time, frequently less than 10%, leading to a tremendous missing data problem. We introduce a principled statistical approach, based on weighted resampling of the observed data, to impute the missing mobility traces, which we then summarize using different mobility measures. We compare the strengths of our approach to linear interpolation, a popular approach for dealing with missing data, both analytically and through simulation of missingness for empirical data. We conclude that our approach offers significant benefits both theoretically and in the actual performance on real-world data.

Keywords: GPS, Imputation, mHealth, Missing data, Mobility

1. Introduction

The Global Positioning System (GPS) is a navigation system that uses a device's distance from a number of satellites in orbit to determine the location of the device. GPS

has a wide range of applications. For example, in transportation GPS has been used to complement self-report surveys on travel activity (Chapman and Frank, 2007; Stopher et al., 2007; Zhou and Golledge, 2003; Shen and Stopher, 2014). One study instrumented participants with both a GPS receiver and an accelerometer and showed that combining both data sources improved the prediction of a person's mode of activity (Troped et al., 2008). Another study outfitted children with GPS devices and showed that their travel behavior was different when they were accompanied by an adult as opposed to when they were on their own (Mackett et al., 2007).

Another burgeoning area of interest is the application of GPS to social and behavioral research (Wolf and Jacobs, 2010). GPS devices on older (> 63 years of age) care-recipients were used to show that caregiver burden was negatively correlated with the amount of time a care-recipient spent walking per day (Werner et al., 2012). Two studies found that mobility measures extracted from GPS data were correlated with depressive symptom severity (Saeb et al., 2015; Canzian and Musolesi, 2015), and another study was able to predict changes of state for bipolar patients with 80.8% accuracy (Gruenerbl et al., 2014). Following spine surgery, GPS tracking was used to monitor patient recovery and found that increased mobility corresponded with a successful recovery (Yair et al., 2011). GPS has also shown promise at keeping track of wandering dementia patients (Miskelly, 2005) as well as at monitoring the mobility of patients with Alzheimer's disease (Shoval et al., 2008). By combining pollution measurements from air samples and a person's GPS trace, exposure levels at the individual-level can be calculated (Phillips et al., 2001).

In most of the above studies, the GPS device was provided to study participants either as a component of a smartphone or as a distinct wearable GPS receiver. While data collected in this way will have minimal missingness, there are three distinct disadvantages to such an approach. Firstly, this study design is not scalable because it is expensive to provide GPS devices to a large number of participants. Secondly, adherence to wearable devices typically declines sharply after a few months (AsPC., 2015) making long-term longitudinal studies less feasible. Thirdly, the data may be biased in unpredictable ways due to the interference that can arise by introducing of a new device into

a participant's life (Ainsworth et al., 2013). All of these shortcomings can be avoided by taking advantage of built-in GPS devices in the smartphones of participants. In 2015, 64% of U.S. adults owned smartphones, up from 35% in 2011 (Smith, 2015), and this number is expected to increase. Anonymized call detail records (CDRs), resulting from mobile phone communication events, have been used to study both social networks (Onnela et al., 2007) and mobility patterns (Gonzalez et al., 2008) at scale, and analysis and modeling of CDRs to different purposes has since then become an active field of its own (Blondel et al., 2015). However, for the purposes of inferring mobility metrics, these data are quite limited as the location of the person is only available at the level of the cell tower used to transmit the event, and even this is only available at the time of communication (either via calls or text messages). Smartphone-based mobility traces from GPS are therefore more precise both spatially and temporally and, importantly, make it possible to link the smartphone data with individual-level covariates, which in the context of digital phenotyping could range from simple demographic variables to fMRI imaging or genome sequencing data.

To make use of the ubiquity of smartphones in biomedical research settings, we have developed a smartphone-based research platform called Beiwe that includes customizable iOS and Android smartphone apps that, among other features, can record a phone's GPS trace using user specified sampling scheme (Torous et al., 2016). For a smartphone app to be scalable and enable long-term data collection, it must not impose too much of a drain on the phone's battery. Study adherence, which in this context means not uninstalling the app during the study, could be in jeopardy if the participant notices a significant drop in the battery life. Of all the current smartphone sensors, GPS is the most expensive with regards to battery usage (Miller, 2012). To lower the strain on the battery, our Beiwe platform records GPS over short intervals, called on-cycles, between gaps of periods of inactivity, called off-cycles. The platform enables users to specify the length of on-cycles and off-cycles at will; for example, the on-period might be 2 minutes and the off-period 10 minutes. However, all of these off-periods lead to a large portion of the GPS trace being missing.

To the best of our knowledge, there is currently no principled method of handling

missing GPS data (Krenn et al., 2011; Jankowska et al., 2015). Studies so far have either ignored missing data (Canzian and Musolesi, 2015), or have used linear interpolation assuming travel at constant 1 m/s over the missing interval (Rhee et al., 2007, 2011). With the rise of research in *digital phenotyping*, which we have previously defined as the “moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices” (Onnela and Rauch, 2016), large scale research studies aiming to measure patient-centered outcomes in naturalistic settings over long periods of time are becoming more prevalent, necessitating the development of statistical methods that properly account for missingness. Here we introduce a statistical approach for imputing the missing trajectories present in a mobility trace that attempts to simulate human mobility patterns. The properties of this approach are compared analytically to linear interpolation, and also to an empirical time series with simulated missingness of a subject for whom high-frequency GPS trajectories were available for one week. Compared to linear interpolation, our approach to imputation offered significant reductions in bias in the estimation of a wide variety of mobility measures, where the average bias of our approach ranged from 83% to 57% of the bias in the estimates produced by linear interpolation.

2. Methods

2.1. Mapping longitude and latitude to a 2D plane

While raw GPS data consists of a sequence of longitude and latitude coordinates that trace a person’s location on the surface of the Earth, most mobility metrics are computed for data in a 2D Euclidean plane, requiring a transformation of coordinates as the first step. Because of the differences in geometry, there will always be distortion when mapping the surface of a 3D sphere to a 2D plane, although the distortion is smaller the smaller the area of the sphere being mapped. People typically do not travel far enough on a daily basis for this projection to greatly distort their mobility trace. For the purposes of extracting mobility measures, a person’s mobility trace can be mapped to a 2D plane on an individual basis. By allowing each person their own projection, distortion can be minimized by selecting the projection best suited for each individual.

Consider a person's mobility trace where we let λ_{min} , λ_{max} , ϕ_{min} , and ϕ_{max} be the minimum and maximum latitude and longitude attained over the study period, respectively. By projecting the region bounded by the points $(\phi_{min}, \lambda_{min})$, $(\phi_{max}, \lambda_{min})$, $(\phi_{min}, \lambda_{max})$, and $(\phi_{max}, \lambda_{max})$ onto an isosceles trapezoid, the distortion of the projection is greatly reduced (see Figure 1). To map a specific (ϕ, λ) coordinate to the X-Y plane, let $w_\lambda = \frac{\lambda - \lambda_{min}}{\lambda_{max} - \lambda_{min}}$, $w_\phi = \frac{\phi - \phi_{min}}{\phi_{max} - \phi_{min}}$, $d_1 = (\lambda_{max} - \lambda_{min}) \cdot R$, $d_2 = (\phi_{max} - \phi_{min}) \cdot R \cdot \sin(\pi/2 - \lambda_{max})$, and $d_3 = (\phi_{max} - \phi_{min}) \cdot R \cdot \sin(\pi/2 - \lambda_{min})$, where $R = 6.371 \cdot 10^6$ m represents the Earth's radius in meters. Then the corresponding (x, y) pair is:

$$\begin{aligned} x &= w_1 \left(\frac{d_3 - d_2}{2} \right) + w_2 \{ d_3(1 - w_1) + d_2 w_1 \} \\ y &= w_1 d_1 \sin \left\{ \cos^{-1} \left(\frac{d_3 - d_1}{2d_1} \right) \right\} \end{aligned}$$

As a reference point, we assign the origin to $(\phi_{min}, \lambda_{min})$.

2.2. Simulating missing trajectories

After converting GPS latitude and longitude to 2D plane coordinates using the above projection, the data remains voluminous, noisy, and difficult to interpret. We therefore use the transformed data to construct meaningful mobility measures that can be used to investigate the mobility patterns and behaviors of study subjects. Once missingness has been accounted for, using the so-called rectangular method of Rhee et al. (2007), the data are converted into a mobility trace defined by a sequence of flights, corresponding to segments of linear movements, and pauses, corresponding to periods of time where a person does not move. Also, if a missing interval is flanked by two pauses at the same location (situated within 50 meters of one another), the missing interval is assumed to be a longer pause at the same location. Our approach to dealing with missing data is to simulate flights and pauses over the period of missingness where the direction, duration, and spatial length of each flight, the fraction of flights versus the fraction of pauses, and the duration of pauses are sampled from observed data. We consider altogether three different approaches that differ in the specifics of what part of observed data that are used for imputation. In our first approach, the missing portion of the data is imputed by borrowing information from flights and pauses that are temporally

nearby. We explore two other alternatives further below.

Let $t_1 < \dots < t_{n+1}$ be a person's event times over the course of their mobility trace, where an event is either a flight, a pause, or a missing interval. Event times occur at the starting and ending times of flights, pauses, and periods of missingness. Let e_j be the event type of event j and it can take one of the following categorical values: *flight*, *pause*, or *missing*. Let t_j be the starting time and t_{j+1} be the ending time of e_j . If e_j is a flight, let a_j be its direction (angle from the eastern axis) and l_j be its distance.

Our goal is to simulate flights and pauses to fill in for each event e_j that is missing. Consider a period of missing data in a mobility trace that starts at t_s and ends at t_f . A person's true Cartesian location is $\mathbf{G}(t) = [G_x(t), G_y(t)]$ and we assume it is unobserved over the time interval (t_s, t_f) with $\mathbf{G}(t_s)$ and $\mathbf{G}(t_f)$ both being known. We aim to closely approximate $\mathbf{G}(t)$ over the missing interval (t_s, t_f) by borrowing information from the observed mobility trace outside this interval. Previous approaches either ignore missing data altogether (Canzian and Musolesi, 2015) or have used linear interpolation between $\mathbf{G}(t_s)$ and $\mathbf{G}(t_f)$, which essentially amounts to connecting the dots at $\mathbf{G}(t_s)$ and $\mathbf{G}(t_f)$ with a straight line. More precisely, linear interpolation assumes that $\mathbf{G}(t) = \frac{t_f - t}{t_f - t_s} \mathbf{G}(t_s) + \frac{t - t_s}{t_f - t_s} \mathbf{G}(t_f)$ on the interval (t_s, t_f) , or the same linear trajectory is assumed at some pre-specified constant velocity, such as 1 m/s, from $\mathbf{G}(t_s)$ to $\mathbf{G}(t_f)$ (Rhee et al., 2007, 2011). While such a simple model of human mobility may be suitable for complete data with scarce missingness, for smartphone GPS data with substantial degree of missingness, a more careful and statistically principled treatment is required.

To simulate a trajectory at any given time point t , we first determine whether a flight or a pause occurs. Two consecutive pauses are not allowed as it should instead just be considered a single longer pause. Therefore if the event prior to t is a pause, then we automatically determine the next event to be a flight. In contrast, if the event prior to t is a flight, then either a flight or a pause can follow, and this is modeled using a Bernoulli distribution with the probability for a flight to occur given by

$$\frac{\sum_{j=2}^n I_{\{e_{j-1}=\text{flight}, e_j=\text{flight}\}} K_{\text{TL}}(t_j : t, t_s, t_f, \nu, c)}{\sum_{j=2}^n \left(I_{\{e_{j-1}=\text{flight}, e_j=\text{flight}\}} + I_{\{e_{j-1}=\text{flight}, e_j=\text{pause}\}} \right) K_{\text{TL}}(t_j : t, t_s, t_f, \nu, c)} \quad (1)$$

where the kernel function that upweights events observed nearby in time (temporally local, or TL) is

$$K_{\text{TL}}(t_j : t, t_s, t_f, \nu, c) = \psi_\nu \left(c \cdot \frac{t - t_j}{t_f - t_s} \right) \quad (2)$$

where $\psi_\nu(\cdot)$ is the density function of a t -distribution with ν degrees of freedom. The effective scale parameter c sets the scale of the distribution, and it can be selected to widen (for smaller c) or narrow (for larger c) the kernel. The $t_f - t_s$ in the denominator of the kernel function in Equation (2) serves to adjust for the duration of the missing period: a longer missing period will rely on a wider range of observed events by increasing the variance of the kernel function.

If the Bernoulli outcome in Equation (1) results in 1, then a flight occurs. In this case an event index is sampled from $\{j \in \{1, \dots, n\} : e_j = \text{flight}\}$ with sampling weights $K_{\text{TL}}(t_j : t, t_s, t_f, \nu, c)$ and the direction a_j , distance l_j , and duration $t_{j+1} - t_j$ at the sampled index define the new flight. If instead the outcome is 0, then a pause occurs. In this case the duration of the pause is sampled from the set of event durations $\{t_{j+1} - t_j : e_j = \text{pause}\}$ with sampling weights $K_{\text{TL}}(t_j : t, t_s, t_f, \nu, c)$. This process of sampling from flights and pauses is repeated with $\Delta^{(m)}$ being the displacement at the m th simulated event and with $a^{(m)}$, $l^{(m)}$, and $t^{(m)}$ being its direction, distance, and duration, respectively. When this event is a pause instead of a flight, $\Delta^{(m)} = [0, 0]$ with $a^{(m)}$ and $d^{(m)}$ are undefined. Starting with $m = 1$, simulated events are taken repeatedly until number of steps q is reached such that $\sum_{m=1}^{q+1} t^{(m)} \geq t_f - t_s$ is true, and which point the process is terminated and $\Delta^{(q)}$ is declared as the final step.

Let $N(t) = \max\{k : t_s + \sum_{m=1}^k t^{(m)} \leq t\}$ be the counting process for the number of simulated events taken by time t and let $\tau(N) = t_s + \sum_{m=1}^N t^{(m)}$ be the time of the N th event in the simulated trajectory. Let $\mathbf{H}_c(t)$ be the simulated trajectory, bridged so it

starts at $\mathbf{G}(t_s)$ and ends at $\mathbf{G}(t_f)$:

$$\mathbf{H}_c(t) = \frac{t_f - t}{t_f - t_s} t_f - t_s \left(\mathbf{G}(t_s) + \sum_{m=1}^{N(t)} \Delta^{(m)} \right) + \frac{t - t_s}{t_f - t_s} \mathbf{G}(t_f) \quad (3)$$

The bridging imposes the constraint that $\mathbf{H}_c(t_s) = \mathbf{G}(t_s)$ and $\mathbf{H}_c(t_f) = \mathbf{G}(t_f)$. The simulated events in $\mathbf{H}_c(t)$ that are flights have a slight curvature induced by the bridging with $\mathbf{G}(t_f)$, so for the flights to retain the property of being straight lines we instead use the slightly modified definition

$$\mathbf{H}(t) = \frac{t - \tau\{N(t)\}}{\tau\{N(t) + 1\} - \tau\{N(t)\}} \mathbf{H}_c[\tau\{N(t)\}] + \frac{\tau\{N(t) + 1\} - t}{\tau\{N(t) + 1\} - \tau\{N(t)\}} \mathbf{H}_c[\tau\{N(t) + 1\}] \quad (4)$$

as the final simulated trajectory. This way $\mathbf{H}(t)$ is composed of pauses and straight flights, while having the same endpoints as the simulated events in $\mathbf{H}_c(t)$.

Above we illustrated our process for simulating a person's trajectory over an interval (t_s, t_f) of missing data. A person's full mobility trace is likely to have multiple missing intervals, and this same approach can be applied equally to each missing interval. Imputing over the gaps in a person's mobility trace in this fashion includes variability: repeated imputations over the same missing intervals are likely to produce variable trajectories. Mobility metrics, such as radius of gyration and distance travelled, will also vary with each imputed mobility trace. As a result, repeated imputations can be used to provide confidence bounds for any mobility metric of interest to account for the uncertainty that results from data missingness and subsequent imputation. After repeating simulating the same trajectories B times and calculating the desired mobility metrics each time, the $\alpha/2 \cdot B$ and the $(1 - \alpha/2) \cdot B$ ordered values form the lower and upper confidence bounds of an α -level confidence interval, respectively.

2.3. Choice of kernel

The kernel in Equation (2) upweights observations that are nearby temporally, but there may be some situations where a different choice of kernel is desirable. If, instead, we

wish to upweight observations that occur nearby *in space* instead of nearby *in time*, then a kernel that upweights geographically local observations can be used:

$$K_{\text{GL}}(t_j : t, t_s, t_f, \nu, c) = \psi_\nu \left(c \cdot \frac{\sqrt{\{G_x(t) - G_x(t_j)\}^2 + \{G_y(t) - G_y(t_j)\}^2}}{z} \right) \quad (5)$$

The constant z in the denominator of Equation (5) is not strictly necessary (the scaling constant c could absorb it), but it allows for a more direct comparison across kernels for the same c . Here we use the *spatial displacement* $z = 50$ (in meters) in the geographically local kernel in a similar fashion we use the *temporal displacement* $t_f - t_s$ (in seconds) in the temporally local kernel in Equation (2).

Circadian rhythm imposes a strong periodicity to human behavior in general, and in some situations it may be desirable to leverage this for data imputation. If a person follows a daily routine, we may want to give higher weight to observations that are multiples of 24 hours apart despite of them occurring on a different day. This may be useful in reconstructing a weekday daily commute, for example. To this end, we propose the geographically local circadian routine kernel:

$$\begin{aligned} K_{\text{GLC}}(t_j : t, t_s, t_f, \nu, c) \\ = K_{\text{GL}}(t_j : t, t_s, t_f, \nu, c) \cdot \psi_\nu \left(c \cdot \frac{\min\{|t - t_j| \pmod{s}, s - |t - t_j| \pmod{s}\}}{t_f - t_s} \right) \end{aligned} \quad (6)$$

The time difference $|t - t_j|$ in Equation (6) is taken modulo $s = 86400$ which is the number of seconds in a 24-hour period.

Regarding terminology in remainder of the paper, when referencing the various approaches for missing data imputation in GPS trajectories, we refer to simulation of trajectories using resampling with a temporally local kernel of Equation (2) as TL; simulation of trajectories using resampling with a geographically local kernel of Equation (5) as GL; and simulation of trajectories using resampling with a geographically local and circadian rhythm kernel of Equation (6) as GLC. Finally, we refer to linear interpolation as LI.

3. Results

3.1. Analytical treatment of the expected gap between a mobility trace and its surrogates

Here we consider a model for a person's mobility trace and then compare analytically the performance of our approach to linear interpolation in their ability to approximate the true trajectory. Consider a mobility trace with no pauses where each flight has the same arbitrary duration of one unit of time where the x displacement is independently distributed from the y displacement of a flight. For the x and y flight displacements, let the expectations be functions of t , $\mu_x(t)$ and $\mu_y(t)$, and let the variances be constant, σ_x^2 and σ_y^2 , respectively. We assume each flight to be independent. Though such stringent independence assumptions would lead unrealistic mobility traces, the analytic results that can be derived based on this model will provide some insight into how the extent of missingness is related to the accuracy of trajectories we simulate over the missing periods.

Let $\mathbf{G}(t) = [G_x(t), G_y(t)]$ be the x and y coordinates of a person's mobility trace at time t . For $t \in (0, T)$ the mobility trace is

$$\mathbf{G}(t) = (t - \lfloor t \rfloor) \Delta_{\lfloor t \rfloor}^{(G)} + \sum_{i \leq \lfloor t \rfloor} \Delta_i^{(G)}$$

where $\Delta_i^{(G)} = [\Delta_{xi}^{(G)}, \Delta_{yi}^{(G)}]$ is a random vector representing the x and y displacement of the i th flight. For convenience we assume $\mathbf{G}(0) = [0, 0]$. If instead of continuous time we consider discrete time $t \in \{0, 1, \dots, T\}$, the above simplifies to

$$\mathbf{G}(t) = \sum_{i=1}^t \Delta_i^{(G)}$$

Though $\mathbf{G}(t)$ represents the actual trajectory, assume that the time period $(0, T)$ represents a period of missingness, and thus a period where $\mathbf{G}(t)$ is unobserved. We use a simplified version of our simulated trajectory $\mathbf{H}(t)$, where we only consider integer-

value t to fill in for this missingness:

$$\mathbf{H}(t) = \sum_{i=1}^t \Delta_i^{(H)} + \frac{t}{T} \left(\mathbf{G}(T) - \sum_{i=1}^T \Delta_i^{(H)} \right). \quad (7)$$

Here $\Delta_i^{(H)} = [\Delta_{xi}^{(H)}, \Delta_{yi}^{(H)}]$ is a random vector representing the x and y displacements of the i th flight, and $\Delta_i^{(H)}$ follows the same distribution as $\Delta_i^{(G)}$, independently. In addition, we have the restriction that $\mathbf{H}(0) = \mathbf{G}(0)$. The second term of Equation (7) bridges $\mathbf{H}(t)$ such that $\mathbf{H}(T) = \mathbf{G}(T)$. Compare this to linear interpolation, which we define as

$$\mathbf{F}(t) = \frac{t}{T} \mathbf{G}(T) \quad (8)$$

Ideally, the simulated trajectory $\mathbf{H}(t)$ and the linearly interpolated trajectory $\mathbf{F}(t)$ should be ‘close’ to the true but unobserved trajectory $\mathbf{G}(t)$ since $\mathbf{H}(t)$ and $\mathbf{F}(t)$ are used as its surrogates. To measure closeness, we examine the average squared distance between the $\mathbf{G}(t)$ and $\mathbf{H}(t)$ across $t \in \{0, 1, \dots, T\}$, $\frac{1}{T+1} \sum_{t=0}^T E \left[\|\mathbf{G}(t) - \mathbf{H}(t)\|^2 \right]$. We then do the same to compare the closeness of $\mathbf{G}(t)$ and $\mathbf{F}(t)$. We seek to answer the question of how the length of the period of missingness, T , relates to the accuracy of the surrogate trajectories $\mathbf{H}(t)$ and $\mathbf{F}(t)$ used to replace the unobserved true trajectory $\mathbf{G}(t)$.

We consider a family of trajectories to allow for varying degrees of curvature in $\mathbf{G}(t)$. For a fixed $\theta_0 \in [0, \pi/2]$ we consider the mean displacements of the flight at time $t \in \{0, 1, \dots, T-1\}$ to be

$$\mu_x(t) = \sqrt{d} \cos \left(\theta_0 - \frac{2\theta_0 t}{T-1} \right) \quad (9)$$

$$\mu_y(t) = \sqrt{d} \sin \left(\theta_0 - \frac{2\theta_0 t}{T-1} \right) \quad (10)$$

where d is the expected distance of a flight. Under this model, $\theta_0 = 0$ corresponds to a straight trajectory whereas $\theta_0 = \pi/2$ corresponds to a semicircular trajectory (see Figure 2). Letting $\bar{D}_x = \frac{1}{T} \sum_{i=1}^T (\Delta_{ix}^{(G)} - \Delta_{ix}^{(H)})$ and $\bar{D}_y = \frac{1}{T} \sum_{i=1}^T (\Delta_{iy}^{(G)} - \Delta_{iy}^{(H)})$, we can

investigate how close the simulated trajectory $\mathbf{H}(t)$ is to the actual trajectory $\mathbf{G}(t)$:

$$\begin{aligned} E \left[\|\mathbf{G}(t) - \mathbf{H}(t)\|^2 \right] &= E \left[\left\| \begin{bmatrix} \sum_{i=1}^t (\Delta_{ix}^{(G)} - \Delta_{ix}^{(H)} - \bar{D}_x) \\ \sum_{i=1}^t (\Delta_{iy}^{(G)} - \Delta_{iy}^{(H)} - \bar{D}_y) \end{bmatrix} \right\|^2 \right] \\ &= 2t(1 - \frac{t}{T})(\sigma_x^2 + \sigma_y^2) \end{aligned}$$

The detailed derivation is in the Supporting materials. By averaging this quantity across all t in the missing interval we arrive at:

$$\begin{aligned} \frac{1}{T+1} \sum_{t=0}^T E \left[\|\mathbf{G}(t) - \mathbf{H}(t)\|^2 \right] &= \frac{1}{T+1} \sum_{t=0}^T 2t(1 - \frac{t}{T})(\sigma_x^2 + \sigma_y^2) \\ &= \frac{2}{T+1} \left\{ \frac{T(T+1)}{2} - \frac{T(T+1)(2T+1)}{6T} \right\} (\sigma_x^2 + \sigma_y^2) \\ &= \left\{ T - \frac{(2T+1)}{3} \right\} (\sigma_x^2 + \sigma_y^2) \\ &= \left(\frac{T-1}{3} \right) (\sigma_x^2 + \sigma_y^2) \end{aligned} \tag{11}$$

Now we look to see how close the linearly interpolated trajectory $\mathbf{F}(t)$ is to $\mathbf{G}(t)$, to compare its performance to that of $\mathbf{H}(t)$. Letting $\bar{\Delta} = \frac{1}{T} \sum_{i=1}^T \Delta_i^{(G)}$:

$$\begin{aligned} E \left[\|\mathbf{G}(t) - \mathbf{F}(t)\|^2 \right] &= E \left[\left\| \begin{bmatrix} \sum_{i=1}^t (\Delta_{ix}^{(G)} - \bar{\Delta}_x) \\ \sum_{i=1}^t (\Delta_{iy}^{(G)} - \bar{\Delta}_y) \end{bmatrix} \right\|^2 \right] \\ &= t \left(1 - \frac{t}{T} \right) (\sigma_x^2 + \sigma_y^2) + M(t) \end{aligned}$$

where

$$\begin{aligned} M(t) = & \sum_{l \in \{x, y\}} \left[\sum_i^t \mu_l(i)^2 + \frac{t^2}{T^2} \sum_{i=1}^T \sum_{j=1}^T \mu_l(i) \mu_l(j) - \frac{2}{T} \sum_{i=1}^t \sum_{j=1}^T \mu_l(i) \mu_l(j) \right. \\ & \left. + 2 \sum_{i < j}^t \left\{ \mu_l(i) \mu_l(j) - \frac{1}{T} \sum_{k=1}^T \mu_l(k) (\mu_l(i) + \mu_l(j)) \right\} \right] \end{aligned}$$

is a function of the $\mu(i)$. The derivation is left for the Supporting materials. Again we average across all time points in the missing interval to arrive at

$$\frac{1}{T+1} \sum_{t=0}^T E \left[\|\mathbf{G}(t) - \mathbf{F}(t)\|^2 \right] = \left(\frac{T-1}{6} \right) (\sigma_x^2 + \sigma_y^2) + \frac{1}{T+1} \sum_{t=0}^T M(t). \quad (12)$$

When comparing the expected gap between $\mathbf{H}(t)$ and $\mathbf{G}(t)$ in Equation (11) and the expected gap between $\mathbf{F}(t)$ and $\mathbf{G}(t)$ in Equation (12), only Equation (12) has both a σ^2 component as well as a component comprised of μ . The σ^2 term in Equation (12) is exactly 1/2 of that in Equation (11), and while the second component involving μ disappears in the case where $t = 0$, $t = T$, or when $\mu(i)$ is constant, but in other cases it can add considerably to the expected gap. Only when $\theta_0 = 0$ is $\mu(i)$ always constant. In this case we would expect the average squared distance between $\mathbf{G}(t)$ and $\mathbf{H}(t)$ to be twice as large as the average squared distance between $\mathbf{G}(t)$ and $\mathbf{F}(t)$. In other words, when expected trajectory has no curvature (i.e., it is a straight line), then linear interpolation is the best approximation of the true trajectory. As the true trajectory gains curvature, $\mathbf{H}(t)$ becomes a closer approximation to the true trajectory and linear interpolation becomes increasingly inaccurate (see Figure 3).

This result tells us that using simulated trajectories from the distribution of *unobserved* flights leads to a better accuracy, on average, than using linear interpolation to fill in the missing data; the only exception to this is if the true unobserved trajectory happens to be a straight line. While this result is demonstrated on a model that assumes we are able to simulate from the distribution of unobserved flights, which is generally not possible since normally only the distribution of *observed flights* is available, our goal is to come as close as possible to this scenario by borrowing information from the ‘closest’ observed flights. As we elaborated above, ‘close’ could mean temporally close (using TL kernel), spatially close (using GL kernel), or close in the sense of leveraging the periodicity of human behavior due to the circadian rhythm along with spatial closeness (GLC kernel).

3.2. *Variability in the biases of mobility measure estimation*

Regardless of the approach used for imputing over the missing intervals in a person's mobility trace, there can still be substantial bias in the mobility estimates that are calculated from the imputed data. After all, each approach assumes a different model; linear interpolation assumes constant linear movement over a missing interval, TL assumes that flights and pauses that occur nearby in time come from the same distribution, GL assumes that flights and pauses that occur nearby in space come from the same distribution, and GLC assumes that flights and pauses that occur at the same time of day and at the same place come from the same distribution. These models each try to approximate the true nature of human mobility, but seldom will any of these models precisely hold true.

In addition, in most cases it is difficult to predict the direction of bias. We demonstrate this through example by looking at one easily interpreted mobility measure: distance travelled. Consider a person who follows a semicircular trajectory with some added jitter to their movement (see Figure 4). Evenly spaced intervals of different sizes are removed to show how the bias changes as the extent of missingness increases/decreases. As expected, in each case as missingness decreases, the bias in the estimates of distance travelled decreases. The bias is predictably negative for LI, and this is intuitive because LI takes the shortest possible path over missing intervals, so it attains the lower bound for the distance travelled metric. In contrast, the TL model is less predictable in the direction of its bias. For the smoother trajectories the TL approach overestimates distance travelled, but for a large enough jitter the bias switches direction and becomes negative. In this small toy example GL will mirror TL in how the data is weighted, and there is no routine for GLC to take advantage of, so both GL and GLC are omitted here.

Overall, the 95% confidence band of TL accurately reflects the amount of missingness in the data, visible as the narrowing of the confidence band as more and more data is observed, whereas the LI approach as a point estimator shows equally misplaced certainty regardless of the amount of missing data. The fact that the confidence bands do not in general attain the nominal 95% coverage of the true distance travelled demonstrate the flaws of the TL assumption that nearby flights come from the same

distribution. Clearly in a semi-circular trajectory this assumption does not hold, and this assumption is not likely to hold in most realistic settings, either. Despite this, it may sometimes be the best choice for filling in the gaps in the data available to us. Unfortunately, when the majority of data is missing, it will be nearly impossible to avoid all bias when estimating various measures of mobility. Instead, one must choose a modeling assumption guided by domain specific knowledge that is as close to the truth as possible. To this end, in the next section, we compare various missing data imputation approaches across many measures of mobility in the context of empirical GPS data.

3.3. Mobility measure estimation on a week-long empirical mobility trace

To generate a high-frequency GPS mobility trace, we had a test subject install an Android version of the Beiwe application on their phone for one week. The application was set to sample the smartphone GPS essentially in continuous time: the on-cycle was specified to be 119 minutes and the off-cycle just 1 minute. Ultimately, due to the occasional loss of power or GPS signal to their phone, an average of 92 minutes of GPS trajectories per day were missing as opposed to the expected 12 minutes per day, but the high quality of this data set allows us to establish a *de facto* ground truth. The goal of this analysis is to take a subset of this data set and to simulate a higher rate of missingness, one that is likely observed in practice. We superimposed on top of the observed data a simulated 2-minute on-cycle and 10-minute off-cycle, and we calculated a variety of mobility measures on the data produced by multiple missing data imputation approaches (LI, TL, GL, and GLC). Here we report the bias for each approach on their estimated mobility measures as compared to the ground truth.

The person's daily mobility trace for the full week is displayed in Figure 5. The mobility trace based on the simulated 2-minute on-cycle and 10-minute off-cycle (top row) is shown alongside the mobility trace based on the complete data (bottom row). The general movements, locations, and daily routines are accurately captured by the subset with missingness, but some of the details are of course lost.

For each day, 15 different mobility measures were calculated (detailed in the Sup-

porting materials), once for each missing data imputation approach and once for the ground truth. The estimates of the mobility measures for one example day are given in Table 1. For each of the resampling approaches, TL, GL, and GLC, three different kernel parameter settings were considered. In each case $\nu = 1$, but the scale parameter was varied ($c = 1, 10, 20$). Increasing the value of c gives greater weight to nearby observations in resampling. The bias is calculated by subtracting the estimated measure under each missing data imputation approach from that same measure calculated on the full data (with near-continuously gathered GPS). For the simulation-based imputation approaches, we used the mean value of the estimated measure from 100 simulated samples in the bias calculations.

A small bias over most of the mobility measures would indicate that the missing data imputation approach does a good job of mimicking real human mobility patterns. To quantify this, each missing data imputation approach is given an overall score between 0 and 1 to provide a means of comparison. These scores are averaged across all days of the week and across all mobility measures. A score of 0 means that there was no bias present for any measure estimation based on that approach, while a score of 1 means that the approach had the largest bias of all the approaches considered for every single measure (Table 2). Based on this metric, the worst performing missing data imputation approach is linear interpolation (LI), with bias that is consistently larger than the resampling-based approaches for the majority of the mobility measures. The best performing missing data imputation approach for this data is GLC with a scaling parameter $c = 10$.

4. Discussion

Past studies with small subject pools have not adequately accounted for missingness, likely because missing data is less of a problem for studies that provide their subjects with dedicated instrumentation capable of recording continuous or near-continuous GPS trajectories. However, instrumenting each subject is expensive and therefore scaling up to larger sample sizes or longer follow-up times becomes infeasible. In the near future, studies will likely increasingly leverage the high ownership rates of smart-

phones so that subjects need only to download an app onto their personal devices. For example, our smartphone research platform Beiwe is currently used to study patient-centered outcomes across different disorders, from depression to surgical recovery, by collecting sensor data, survey data, and phone usage patterns from diverse patient cohorts. In these studies, which generally have long follow-up times, battery life is preserved by recording GPS less frequently, often leading to more than 80% missing data. This means that missingness can no longer be ignored and will need to be properly adjusted for. In this paper, we introduced a new statistical approach to address missingness, and we found that, even with large percentages of missingness, mobility measure estimation that uses the proposed data imputation is accurate compared the current standard of using linear interpolation.

With the prospect of scalable studies on the horizon, additional statistical challenges will likely emerge in the analysis of mobility measures from patient cohorts. With mobility measures paired with daily smartphone surveys, the longitudinal nature of the data can be leveraged with generalized linear mixed models (Breslow and Clayton, 1993) (GLMM) or generalized estimating equations (Liang and Zeger, 1986) (GEE) to estimate the effects of mobility measures on various outcomes obtained through the surveys. Also, while here we considered only mobility measures extracted from GPS traces, these mixed model frameworks can be readily adapted to include information from other smartphone sensors by adding additional covariates, such as those obtained from the phone's built-in accelerometer, into the regression model.

Finally, the method introduced in this paper has been implemented as a package in the statistical computing software, R, and is freely available (see Supporting materials).

References

Ainsworth, J., Palmier-Claus, J. E., Machin, M., Barrowclough, C., Dunn, G., Rogers, A., Buchan, I., Barkus, E., Kapur, S., Wykes, T. et al. (2013) A comparison of two delivery modalities of a mobile phone-based assessment for serious mental illness: native smartphone application vs text-messaging only implementations. *Journal of medical Internet research*, **15**, e60.

- AsPC., I. (2015) Building a better tracker: Older consumers weigh in on activity and sleep monitoring devices.
- Blondel, V. D., Decuyper, A. and Krings, G. (2015) A survey of results on mobile phone datasets analysis. *arXiv preprint arXiv:1502.03406*.
- Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, **88**, 9–25.
- Canzian, L. and Musolesi, M. (2015) Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 1293–1304. ACM.
- Chapman, J. and Frank, L. (2007) Integrating travel behavior and urban form data to address transportation and air quality problems in atlanta.
- Gonzalez, M. C., Hidalgo, C. A. and Barabasi, A.-L. (2008) Understanding individual human mobility patterns. *Nature*, **453**, 779–782.
- Gruenerbl, A., Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., Mayora, O., Haring, C. and Lukowicz, P. (2014) Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In *Proceedings of the 5th Augmented Human International Conference*, 38. ACM.
- Jankowska, M. M., Schipperijn, J. and Kerr, J. (2015) A framework for using gps data in physical activity and sedentary behavior studies. *Exercise and sport sciences reviews*, **43**, 48–56.
- Krenn, P. J., Titze, S., Oja, P., Jones, A. and Ogilvie, D. (2011) Use of global positioning systems to study physical activity and the environment: a systematic review. *American Journal of Preventive Medicine*, **41**, 508–515.
- Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

- Mackett, R., Brown, B., Gong, Y., Kitazawa, K. and Paskins, J. (2007) Setting children free: childrens independent movement in the local environment.
- Miller, G. (2012) The smartphone psychology manifesto. *Perspectives on Psychological Science*, **7**, 221–237.
- Miskelly, F. (2005) Electronic tracking of patients with dementia and wandering using mobile phone technology. *Age and ageing*, **34**, 497–498.
- Onnela, J.-P. and Rauch, S. L. (2016) Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*.
- Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J. and Barabási, A.-L. (2007) Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, **104**, 7332–7336.
- Phillips, M. L., Hall, T. A., Esmen, N. A., Lynch, R. and Johnson, D. L. (2001) Use of global positioning system technology to track subject's location during environmental exposure sampling. *Journal of exposure analysis and environmental epidemiology*, **11**, 207–215.
- Rhee, I., Shin, M., Hong, S., Lee, K. and Chong, S. (2007) Human mobility patterns and their impact on routing in human-driven mobile networks. In *Proceedings of Hotnets-VI*.
- Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. J. and Chong, S. (2011) On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)*, **19**, 630–643.
- Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P. and Mohr, D. C. (2015) Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research*, **17**.
- Shen, L. and Stopher, P. R. (2014) Review of gps travel survey and gps data-processing methods. *Transport Reviews*, **34**, 316–334.
- Shoval, N., Auslander, G. K., Freytag, T., Landau, R., Oswald, F., Seidl, U., Wahl, H.-W., Werner, S. and Heinik, J. (2008) The use of advanced tracking technologies for

- the analysis of mobility in alzheimer's disease and related cognitive diseases. *BMC geriatrics*, **8**, 1.
- Smith, A. (2015) Us smartphone use in 2015. *Pew Research Center*, 18–29.
- Stopher, P., FitzGerald, C. and Xu, M. (2007) Assessing the accuracy of the sydney household travel survey with gps. *Transportation*, **34**, 723–741.
- Torous, J., Kiang, M., Lorme, J. and Onnela, J.-P. (2016) New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smart-phone research. *JMIR Mental Health*.
- Troped, P. J., Oliveira, M. S., Matthews, C. E., Cromley, E. K., Melly, S. J. and Craig, B. A. (2008) Prediction of activity mode with global positioning system and accelerometer data. *Medicine and science in sports and exercise*, **40**, 972–978.
- Werner, S., Auslander, G. K., Shoval, N., Gitlitz, T., Landau, R. and Heinik, J. (2012) Caregiving burden and out-of-home mobility of cognitively impaired care-recipients based on gps tracking. *International Psychogeriatrics*, **24**, 1836–1845.
- Wolf, P. S. and Jacobs, W. J. (2010) Gps technology and human psychological research: A methodological proposal. *Journal of Methods and Measurement in the Social Sciences*, **1**, 1–7.
- Yair, B., Noam, S., Meir, L., Gail, A., Amit, B., Michal, I., Vaccaro, A. R. and Leon, K. (2011) Assessing the outcomes of spine surgery using global positioning systems. *Spine*, **36**, E263–E267.
- Zhou, J. J. and Golledge, R. (2003) An analysis of variability of travel behavior within one-week period based on gps. *University of California Transportation Center*.

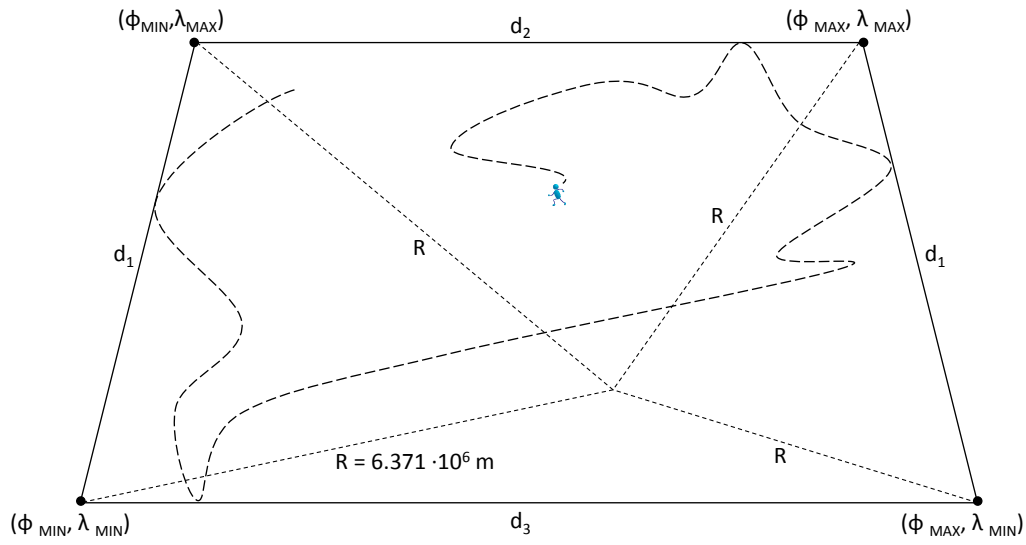


Fig. 1. Schematic of longitude-latitude projection to X-Y plane. The isosceles trapezoid contains the projection of the mobility trace for a particular individual. In the northern hemisphere $d_2 < d_3$ while in the southern hemisphere $d_2 > d_3$. The long dashed curve represents a person's example mobility trace.

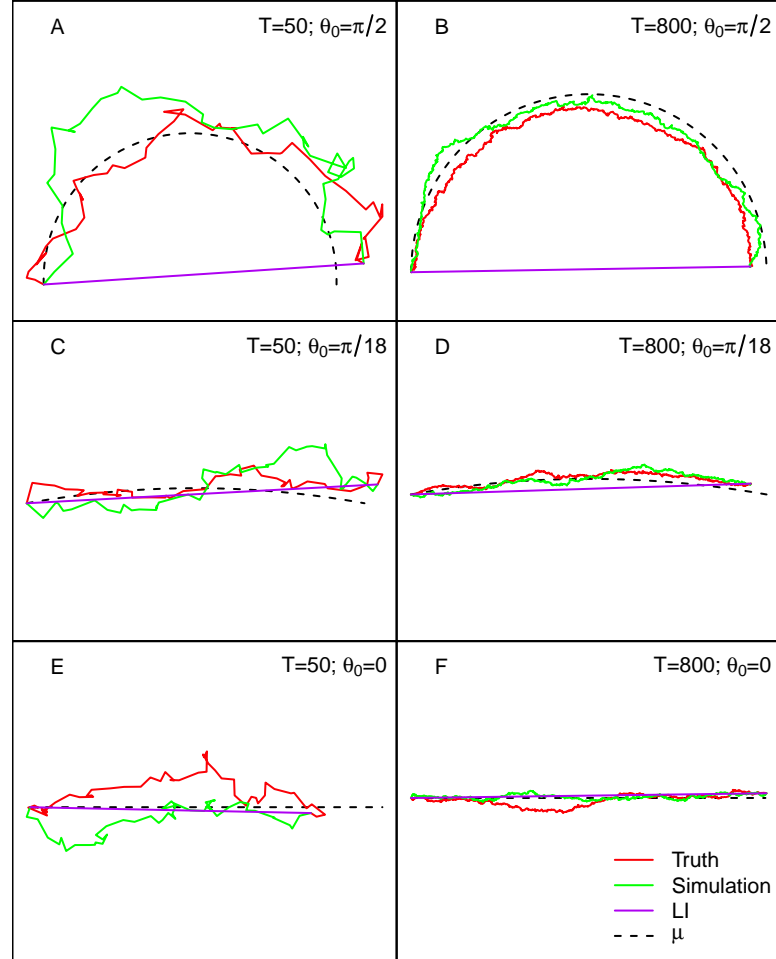


Fig. 2. Theoretical unobserved trajectories and their surrogates. Trajectories are generated according to the theoretical model of Section 3.1. Panels (A), (C), and (E) represent a shorter period of missingness ($T = 50$) while panels (B), (D), and (F) represent a longer period of missingness ($T = 800$). The red trajectory represents a person's true, unobserved mobility trace over an interval of T units of time, and the dashed line represents its expected trajectory. It is assumed that the location immediately before and immediately after this interval are observed and known. The green trajectory represents one simulated instance, while the purple trajectory represents linear interpolation as a means for imputing the missing gap. The θ_0 represents the starting angle above the x -axis of the mean trajectory. Linear interpolation is best when the expected trajectory is a straight line, but only the simulation approach is robust to curvature.

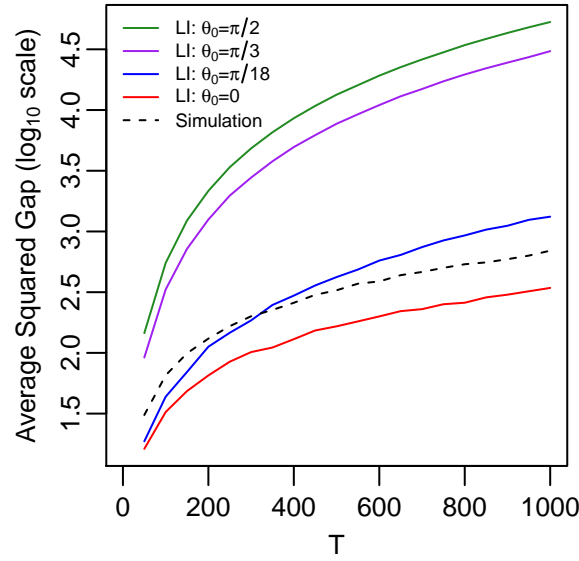


Fig. 3. Expected average gap between imputed trajectories and the true unobserved trajectory. The length of the interval of missingness, T , ranges from 50 to 1000 by increments of 50. For each value of T , the average over 1000 simulations are used to estimate the average squared gap for the simulation approach. While for small T and small θ_0 linear interpolation can be a better approximation to the true trajectory, asymptotically the simulation approach is better for any amount of curvature.

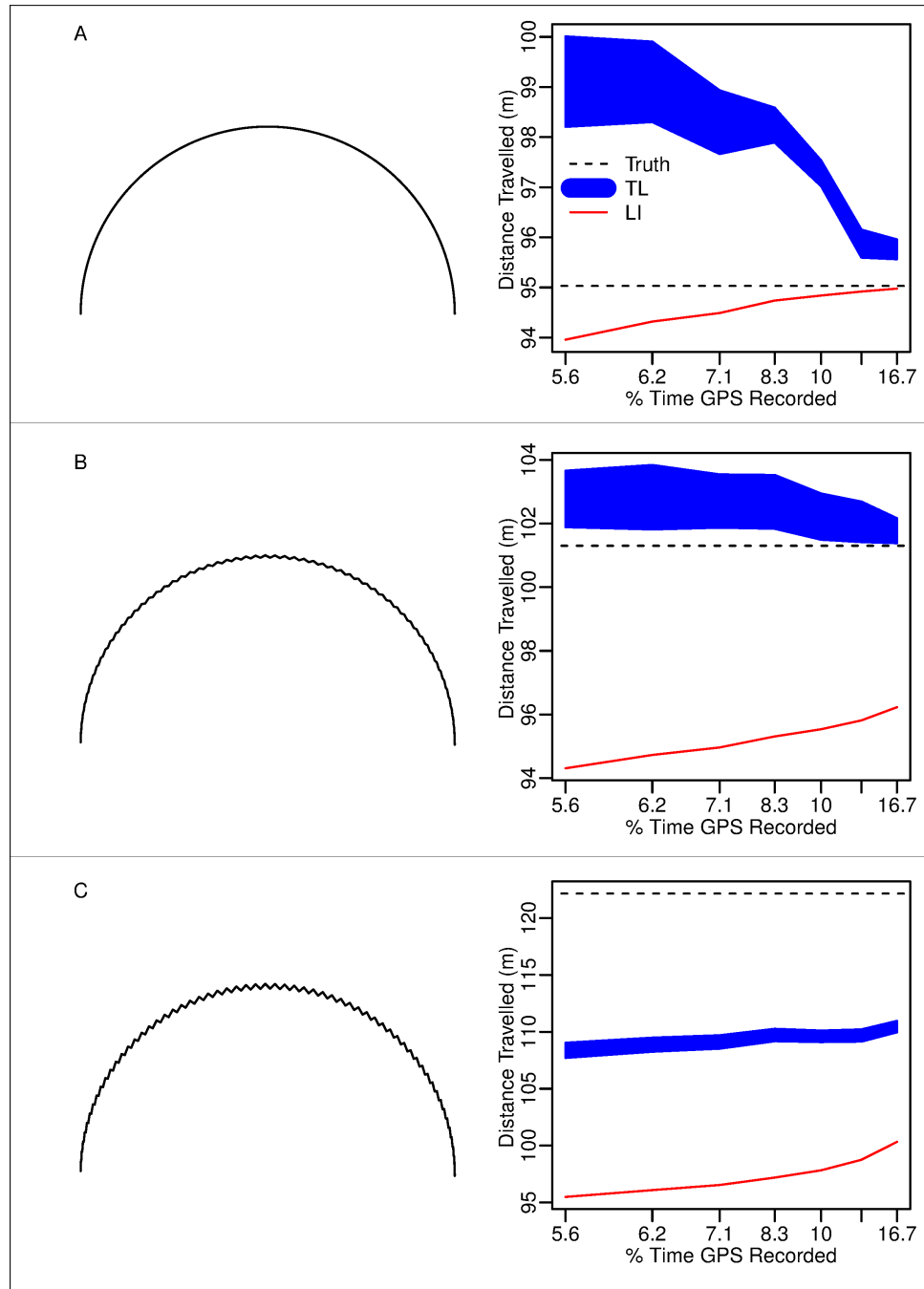


Fig. 4. Bias in the estimation of mobility metrics as the amount of missingness changes.

On the left, three different trajectories are displayed. The right of each panel shows the estimates of distance travelled for various levels of missingness in the trajectory on the left. Missingness is generated by taking evenly spaced intervals of different sizes (for different levels of missingness) out of the semicircular trajectories. For TL, each level of missingness is repeated for 100 simulated trajectories to obtain the 95% confidence band. This is repeated for three different types of movement: a smooth trajectory (A), small jitters (B), and larger jitters (C). This demonstrates that both the direction and magnitude of a surrogate's bias in approximating the true trajectory can vary significantly depending on the true trajectory.

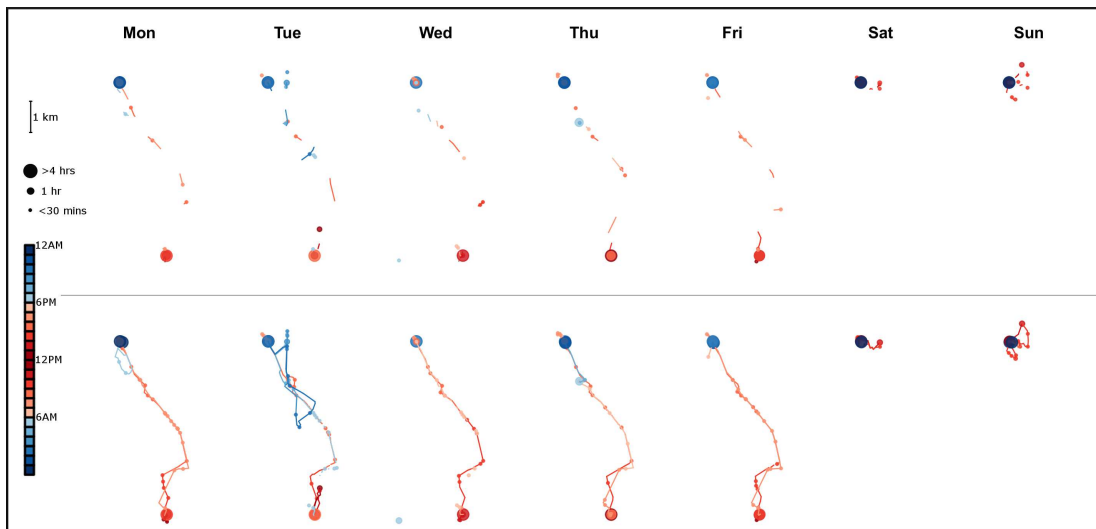


Fig. 5. A person's example daily trajectories over the course of a week. The bottom row represents a person's trajectory when GPS is captured continuously. The top row represents the identical trajectories to the bottom row except with GPS assumed to be recorded only for two minute intervals with ten minute gaps of missingness between recorded intervals. Lines represent flights, or movement. Points represent pauses, or periods where the person is stationary, with larger points indicating longer pauses.

Table 1. Example of a person's mobility measures compared across different missing data imputation approaches. GPS was collected continuously to establish the ground truth (92.17 minutes are missing due to either loss of signal or power). For the missing data imputations, a Cauchy kernel was used with scale factor denoted by the number following the period. Larger scale factors give increased weight on nearby observations during resampling. For the TL, GL, and GLC approaches, the margin of error represents the standard deviation over 100 repeated simulations.

Measures	TL.1	TL.10	TL.20	GL.1	GL.10	GL.20	GLC.1	GLC.10	GLC.20	LI	Truth
Hometime	831.5 ±2.3	832.3 ±2.4	833.4 ±2.2	830.3 ±2.2	830.5 ±2.8	829.8 ±1.9	829.1 ±2.1	832.1 ±2.2	831.3 ±2.5	826.7	882.8
DistTravelled	22184 ±969.7	22446 ±843.5	22569 ±811.6	18801 ±466.3	18801 ±337.5	18779 ±369.4	21791 ±969.9	22380 ±712.1	22444 ±645.6	17236	19344
RoG	2787.3 ±2.3	2791.3 ±2.6	2791.2 ±1.9	2783.0 ±1.6	2783.0 ±1.9	2783.3 ±2.5	2785.6 ±1.3	2787.0 ±1.5	2787.5 ±1.8	2779.4	2781.3
MaxDiam	6717 ±169	6745 ±129	6727 ±98	6494 ±44	6483 ±8	6496 ±34	6516 ±55	6517 ±55	6562 ±94	6479	6467
MaxHomeDist	6372 ±165	6410 ±123	6379 ±93	6160 ±49	6147 ±16	6153 ±39	6144 ±30	6152 ±5	6163 ±24	6149	6129
SigLocsVisited	2.96 ±0.73	3.20 ±0.58	3.20 ±0.71	3.16 ±0.69	3.00 ±0.76	2.96 ±0.79	3.28 ±0.61	3.12 ±0.60	3.20 ±0.65	2	3
AvgFlightLen	172.7 ±10.7	160.2 ±7.6	158.6 ±7.4	200.2 ±23.2	193.2 ±19.2	191.7 ±18.1	129.9 ±13.6	122.8 ±6.1	127.1 ±7.6	478.8	251.2
StdFlightLen	152.9 ±30.8	125.8 ±10.1	123.2 ±5.5	213.4 ±51.5	205.8 ±36.3	202.7 ±43.5	151.0 ±30.0	134.2 ±8.4	137.1 ±9.0	639.6	223.3
AvgFlightDur	79.0 ±9.3	69.4 ±5.8	68.8 ±5.6	119.0 ±17.9	115.2 ±13.4	113.5 ±13.7	65.4 ±10.5	57.2 ±4.1	60.0 ±5.1	340.6	77.0
StdFlightDur	131.7 ±17.0	115.3 ±9.0	113.5 ±10.2	170.3 ±22.0	168.7 ±14.8	166.7 ±14.4	103.7 ±18.2	85.0 ±10.9	91.7 ±13.1	289.8	55.2
FracPause	0.88 ±0.01	0.89 ±0.01	0.89 ±0.01	0.87 ±0.01	0.87 ±0.01	0.87 ±0.01	0.87 ±0.01	0.88 ±0.01	0.88 ±0.01	0.86	0.93
SigLocEntropy	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63 ±0.01	0.63	0.63
MinsMissing	1243	1243	1243	1243	1243	1243	1243	1243	1243	1243	92
CircdnRtn	0.64 ±0.02	0.63 ±0.01	0.63 ±0.02	0.67 ±0.01	0.67 ±0.01	0.67 ±0.01	0.65 ±0.02	0.66 ±0.01	0.66 ±0.02	0.69	0.66
WkEndDayRtn	0.76 ±0.02	0.76 ±0.01	0.76 ±0.01	0.78 ±0.01	0.77 ±0.01	0.78 ±0.01	0.76 ±0.02	0.76 ±0.01	0.77 ±0.01	0.81	0.79

Table 2. Comparison of different missing data imputation approaches to the ground truth.

For each measure, the absolute error relative to the ground truth measure is scaled so that the approach with the highest error is set to 1.00. For the stochastic approaches, TL, GL, and GLC, the mean daily measures from 100 simulations are used in relative error calculations. These scores are calculated for each day and averaged across a full week. In this way, a score of 0 implies the best case where the missing data imputation approach matched the ground truth exactly, and a score of 1.00 represents the worst case. The final row is the average score across all measures. The best performing missing data approach, GLC with a Cauchy kernel and scaling parameter of 10, is highlighted in blue. Highlighted in red, linear interpolation performed the worst.

Measures	TL.1	TL.10	TL.20	GL.1	GL.10	GL.20	GLC.1	GLC.10	GLC.20	LI
Hometime	0.90	0.88	0.89	0.95	0.95	0.94	0.96	0.95	0.95	1.00
DistTravelled	0.53	0.57	0.58	0.23	0.31	0.28	0.05	0.22	0.22	1.00
RoG	0.44	1.00	0.87	0.41	0.39	0.53	0.86	0.76	0.74	0.74
MaxDiam	0.98	1.00	0.87	0.25	0.16	0.23	0.12	0.13	0.19	0.06
MaxHomeDist	0.92	1.00	0.79	0.35	0.26	0.32	0.09	0.18	0.19	0.06
SigLocsVisited	0.85	0.95	0.93	0.95	0.87	1.00	0.97	0.90	0.92	0.42
AvgFlightLen	0.30	0.35	0.36	0.17	0.20	0.19	0.39	0.42	0.41	1.00
StdFlightLen	0.07	0.12	0.13	0.04	0.03	0.04	0.11	0.13	0.13	1.00
AvgFlightDur	0.01	0.02	0.02	0.12	0.11	0.12	0.01	0.03	0.03	1.00
StdFlightDur	0.24	0.17	0.18	0.49	0.49	0.50	0.29	0.24	0.25	1.00
FracPause	0.84	0.81	0.81	0.86	0.84	0.84	0.84	0.81	0.81	1.00
SigLocEntropy	0.31	0.39	0.62	0.62	0.38	0.23	0.69	0.38	0.85	1.00
CircdnRtn	0.38	0.31	0.25	0.13	0.13	0.13	0.00	0.25	0.31	1.00
WkEndDayRtn	1.00	0.73	0.67	0.73	1.00	0.67	0.53	0.33	0.13	0.40
Average Score	0.55	0.59	0.57	0.45	0.44	0.43	0.42	0.41	0.44	0.76